Intuit Quickbooks Upgrade

This case is a "classic" in more ways than one. The key event, the release of version 3 of the QuickBooks software, takes place in 1995. Although ecommerce was already feasible in 1991, at the time of the case, QuickBooks products could only be purchased through Intuit Direct (i.e., delivery by mail) or through "brick-and-mortar" retailers like Best Buy.

The purpose of this exercise is to gain experience modeling the response to an upsell campaign. The intuit75.parquet file contains data on 75,000 (small) businesses selected randomly from the 801,821 that were sent the wave-1 mailing.[1] The mailing contained an offer to upgrade to the latest version of the QuickBooks software.

Variable "res1" denotes which of these businesses responded to the mailing by purchasing QuickBooks version 3.0 from Intuit Direct. Note that Intuit Direct sells products directly to its customers rather than through a brick-and-mortar retailer like Best Buy. Use the available data to predict which businesses that did not respond to the wave-1 mailing, are most likely to respond to the wave-2 mailing. Note that variables were added, deleted, and recoded so ***please ignore the variable descriptions in Exhibit 3 in the case on Study.Net.*** Instead, use the variable descriptions in the intuit75k.parquet file and in the table below:

Variable	Type	Description
id	integer	Small business customer ID
zip5	character	5-Digit ZIP Code (00000 = unknown, 99999 = international ZIPs).
zip_bins	integer	Zip-code bins (20 approximately equal sized bins from lowest to highest zip code number)
sex	factor	"Female", "Male", or "Unknown."
bizflag	integer	Business Flag. Address contains a business name (1=yes, 0=no or unknown).
numords	integer	Number of orders from Intuit Direct in the previous 36 months
dollars	numeric	Total $ ordered from Intuit Direct in the previous 36 months
last	integer	Time (in months) since last order from Intuit Direct in the previous 36 months
sincepurch	integer	Time (in months) since original (not upgrade) Quickbooks purchase
version1	integer	Is 1 if the customer's current Quickbooks is version 1, 0 if version 2
owntaxprod	integer	Is 1 if the customer purchased tax software, 0 otherwise
upgraded	integer	Is 1 if customer upgraded from Quickbooks version 1 to version 2
res1	factor	Response to wave-1 mailing ("Yes" if responded else "No")
training	integer	70/30 split, 1 for training sample, 0 for test set

[1] The available data was disguised and altered for pedagogical purposes.

Assignment guidelines:

1. Determine which of the 22,500 businesses in the test set (i.e., training == 0) to mail in wave-2 (i.e., generate a list of IDs). One quarter of the grade (10 points) for this group assignment will be based on the **profit** (*not* ROME) achieved using this list of IDs. After you submit the list of IDs I will be able to determine the final outcome because all customers who did not respond in wave-1 were actually mailed in wave-2. You will **not** have access to this extra data set with information on response in wave-2.

   Your selection of model types should include **Logistic regression** and **Neural Networks**.

2. In your write-up, please scale the profit estimate derived from your best model's performance in the **test** set to the full set of businesses to target in wave-2. Note that of the 801,821 businesses in the wave-1 mailing 38,487 already responded and should not be mailed again.

3. For the purposes of this exercise assume each mail piece costs $1.41 and that the margin (or net revenue) from each responder, excluding the mailing cost, is $60. ***Please ignore numbers in the case on Study.Net that relate to profits and costs***.

4. Please note the following statement in the case in the course reader (page 4): "Usual practice would be to assume a 50 percent drop off in response from wave-1 to wave-2." For your analysis, this means that when you decide whom to mail in wave-2, you should assume that every response probability in wave-2 is only 50% of the response probability you predict for that business based on the wave-1 response data.

5. Fifty percent of the grade (20 points) for this assignment will be based on a report that should explain how your group determined the list of IDs to target. The text in the report, excluding exhibits, should not be more than the equivalent of 3 single-spaced pages (i.e., approx. 1500 words). The report must be completely reproducible and in Jupyter Notebook format. The assignment write-up should be submitted through **GitHub** and then connected to **GradeScope** before 9am on the day of class.

   Answer the following questions in your write-up:

   - Describe how you developed your predictive models, and discuss predictive performance for each model
   - How did you compare and evaluate different models?
   - If you created new variables to include in the model, please describe these as well
   - What criteria did you use to decide which customers should receive the wave-2 mailing?
   - How much profit do you anticipate from the wave-2 mailing?
   - What did you learn about the type of businesses that are likely to upgrade?

6. Your team should create a short presentation video and submit it through Canvas. Instructions are available in "Group Assignment: Intuit Quickbooks Upgrade (Video submission)" on the Canvas assignment page.

7. The presentation should not last more than 5-10 minutes and should cover your approach to solving the case and your key results. Peer evaluation will used, in part, to evaluate the presentation (10 points).

8. **Before 9pm on the day before class** please post the list of IDs you want to mail to Canvas using **"**Group Assignment: Intuit Quickbooks Upload Target IDs". The assignment write-up should be submitted through **GitHub** and then connected to **GradeScope** before 9am on the day of class. The reason that the list of IDs must be submitted earlier is that I need to compile all the results before

class. Compiling the results takes some time so please help me out and stick to the following instructions **exactly**!

Please post a CSV file with **only 2 variables**

a. The original **id** variable from the intuit75.parquet dataset (please do **not** rename the variable). The column should contain all IDs from the test set. Do **NOT** delete IDs you don't want to target from the dataset, i.e., **the file must have 22,500 rows**.
b. A variable named **"mailto_wave2"** (lower case) that is True if you want to target a customer in wave-2, and False if you do not.
c. Please name the dataset using the **first** names of **all** group members separated by underscores "_", plus your group name (e.g., Nancy_Yu_Manuel_MightyDucks.csv). If your group name is more than one word please leave out spaces or underscores (e.g., instead of "Mighty Ducks" use "MightyDucks"). Also, the name should not contain any symbols.
d. There is an example file on Canvas and in the assignment repo on GitHub so you can see the required format
e. **Please double and triple check that your file is in the exact right format, so you do not end up being the group that crashed my code!** 😊

9. Generative AI (5 points): Please describe in detail how your team used Generative AI-tools like ChatGPT to support your work on this case. Provide pdfs and/or screenshots of your "discussions" with these tools and comment on what things did and did not go well. Make sure to add discussion about your thought process and how you tried to maximize the benefits from using these tools. Also add any questions you may have about the assignment and the support you received from GenAI so we can discuss these topics in class.

   Note: No matter how you used Generative AI-tools, you are expected to fully understand all elements of the case solution submitted by your group. Any group member may be called on in class to walk us through your thought process and how different parts of your code work and how you arrived at your solution.

Hints:

- This case is a bit of a "journey of discovery". An important part of that journey relates to the zip-code information. I created zip code bins ("zip_bins") for you labeled 1-20. Each bin contains about 3,750 customers sorted by zip-code
- Efficiently estimating and comparing various models is a critical skill to develop! You can use your own functions or functions from python packages such as pyrsm to calculate model performance metrics (e.g., profit, ROME, AUC, lift, gains, etc.) for various models.
- For the source code of the relevant python functions from the pyrsm package that you can use see: https://github.com/vnijs/pyrsm/blob/main/pyrsm/model/perf.py
- If you want to try running a logistic regression model with variable selection, I recommend using L1 regularization (Lasso) in sklearn. After variable selection, re-estimate a logistic regression model with just the variables you want to keep in your models using pyrsm. Note that there are some important down-sides to this variable selection approach that we will discuss in class.
- Please post questions on Piazza. Choose a 'private' post only if needed.